Units of measure identification in unstructured scientific documents in microbial risk in food

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie-Barthelemy, Mathieu Roche

Units of measure identification in unstructured scientific documents in microbial risk in food

Soumia Lilia Berrahou[1,2], Patrice Buche[1,2], Juliette Dibie-Barthélemy[3] and Mathieu Roche[1]

1. *LIRMM-CNRS - Univ. Montpellier 2, F-34095 Montpellier Cedex 5, France*
2. *INRA - UMR IATE, 2, place Pierre Viala, F-34060 Montpellier Cedex 2 France*
3. *INRA - Mét@risk & AgroParisTech, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France*

OBJECTIVE(S)

A preliminary step in microbial risk assessment in food is to gather and capitalize experimental data. Data capitalization is a crucial stake in an overall decision support system which consists of predicting microbial behavior [1]. In the framework of the French ANR project MAP'OPT (Equilibrium Gas Composition in Modified Atmosphere Packaging and Food Quality), the predictive modeling platform Sym'Previus (www.symprevius.org) should be able to propose a global approach to establish a scientifically sound method for choosing an appropriate modified atmosphere and associated packaging solution.

Our work is part of this overall system and aims at extracting semi-automatically experimental data from unstructured scientific documents. Indeed, these documents use natural language combined with domain-specific terminology that is extremely time-consuming and tedious to extract in the free form of text and therefore to gather and capitalize. Our work relies on the MAP'OPT-Onto ontology [4], which has been built as an extension of the ontology used in Sym'Previus by adding concepts about food packaging, quantity concepts and concepts managing units of measures.

Experimental data are often expressed with concepts (e.g *packaging*, *permeability*) or a numerical value often followed with its unit of measure (e.g. *258 amol m-1 s-1 Pa-1*). In this paper, our work deals with unit recognition, known as a scientific challenge.

METHOD(S)

Extracting automatically quantitative data is a painstaking process because units suffer from different ways of writing within documents. We can encounter same units written in different manners such as *amol* m^{-1} s^{-1} Pa^{-1} written as *amol*.m^{-1}.s^{-1}.Pa^{-1} or as *amol/m/s/Pa*. We aim at focusing on the extraction and identification of these variant units seen as synonyms, in order to enrich iteratively an ontology, which represents a predefined vocabulary used to annotate, capitalize and query experimental data extracted from texts [2]. Our work addresses unit extraction and identification issues from texts to enrich an ontology in a two-step approach. First, we use text-mining methods and supervised learning approaches in order to predict relevant parts of the text where synonyms of units or new units are. The second step of our method consists in extracting specific strings representing units in the segments of texts found in the previous step. The extracted candidates are compared to units already present in the ontology using a new edit measure based on Damerau-Levenshtein [3].

RESULTS

We have made experiments on 115 scientific documents (i.e. around 35 000 sentences) on food packaging. Each unit is recognized from a list of 211 units already defined in the MAP'OPT-Onto. Our learning algorithms predict that almost 5 000 sentences contain units. This prediction is correct for 95,5% of cases. In the second step, we have successfully extracted 38 terms as either synonyms or new units from sentences selected in the first step. So, we can propose 18% of enrichment of the pre-existing MAP'OPT-Onto.

CONCLUSIONS AND IMPACT OF THE STUDY

We propose a two-step approach to enrich an ontology with unit synonyms. Our approach addresses both issues: location and extraction of units. Future work should be defined in order to automatically populate the ontology with new concepts (e.g. food product or packaging names) and link the new units discovered.

REFERENCES

1. *Leporq B., Membré J.-M., Dervin C., Buche P. and Guyonnet J.P, (2005). The ``Sym'Previus'' software, a tool to support decisions to the foodstuff safety. International Journal of Food Microbiology, 100(1-3) 231-237*

2. *Buche P., Couvert O., Dibie-Barthélemy J.,Hignette G., Mettler E., Soler L. (2011). Flexible Querying Of Web Data To Simulate Bacterial Growth In Food. Food Microbiology 28(4):685-693*

3. *Buche P., Dibie-Barthélemy J., Ibanescu L., SolerL. (2013). Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource. IEEE Trans. Knowl. Data Eng. 25 (4): 805-819*

4. *Damerau, F.J.: A technique for computer detection and correction of spelling errors. Commun. ACM 7(3) (1964) 171-176*

5. *Touhami R., Buche P., Dibie-Barthélemy J. and Ibanescu L. (2011). An ontological and terminological resource for n-ary relation annotation in web data tables, in On the Move to Meaningful Internet Systems: OTM 2011, Springer, p. 662–679.*